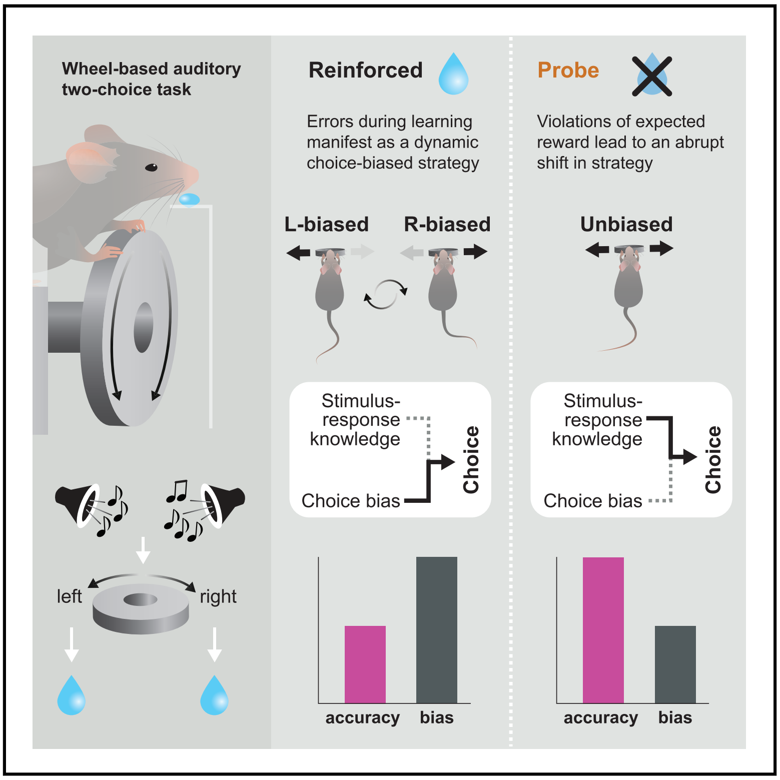
Current Biology

Article

Performance errors during rodent learning reflect a dynamic choice strategy

Graphical abstract


Wheel-based auditory two-choice task
left
right
Reinforced
Errors during learning manifest as a dynamic choice-biased strategy
L-biased
R-biased
Stimulus-response knowledge
Choice bias
Choice
accuracy
bias
Probe
Violations of expected reward lead to an abrupt shift in strategy
Unbiased
Stimulus-response knowledge
Choice bias
Choice
accuracy
bias

Highlights

- Errors while learning a two-choice task are dominated by intentional strategies
- Strategies manifest as a dynamic choice bias, testing one option or another in bouts
- Mice exhibit sensitivity to violations of expected reward during biased epochs
- Violating expectations leads to an abrupt shift in strategy

Authors

Ziyi Zhu, Kishore V. Kuchibhotla

Correspondence

kkuchib1@jhu.edu

In brief

Errors during learning are typically considered mistakes. Zhu and Kuchibhotla challenge this dogma, demonstrating that many errors made by rodents during learning are purposeful, reflecting a structured form of exploration. They combine quantitative behavior, kinematic analysis of behavioral choices, and computational modeling to support their findings.

Zhu & Kuchibhotla, 2024, Current Biology *34*, 2107–2117
May 20, 2024 © 2024 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.cub.2024.04.017

CellPress

Article

Performance errors during rodent learning reflect a dynamic choice strategy

Ziyi Zhu[1,2,3] and Kishore V. Kuchibhotla[1,2,3,4,5,6,*]
[1]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21218, USA
[2]Johns Hopkins Kavli Neuroscience Discovery Institute, Johns Hopkins University, Baltimore, MD 21218, USA
[3]The Solomon Snyder Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[4]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[5]X (formerly Twitter): @kishoreneuro
[6]Lead contact
*Correspondence: kkuchib1@jhu.edu

https://doi.org/10.1016/j.cub.2024.04.017

SUMMARY

Humans, even as infants, use cognitive strategies, such as exploration and hypothesis testing, to learn about causal interactions in the environment. In animal learning studies, however, it is challenging to disentangle higher-order behavioral strategies from errors arising from imperfect task knowledge or inherent biases. Here, we trained head-fixed mice on a wheel-based auditory two-choice task and exploited the intra- and inter-animal variability to understand the drivers of errors during learning. During learning, performance errors are dominated by a choice bias, which, despite appearing maladaptive, reflects a dynamic strategy. Early in learning, mice develop an internal model of the task contingencies such that violating their expectation of reward on correct trials (by using short blocks of non-rewarded "probe" trials) leads to an abrupt shift in strategy. During the probe block, mice behave more accurately with less bias, thereby using their learned stimulus-action knowledge to test whether the outcome contingencies have changed. Despite having this knowledge, mice continued to exhibit a strong choice bias during reinforced trials. This choice bias operates on a timescale of tens to hundreds of trials with a dynamic structure, shifting between left, right, and unbiased epochs. Biased epochs also coincided with faster motor kinematics. Although bias decreased across learning, expert mice continued to exhibit short bouts of biased choices interspersed with longer bouts of unbiased choices and higher performance. These findings collectively suggest that during learning, rodents actively probe their environment in a structured manner to refine their decision-making and maintain long-term flexibility.

INTRODUCTION

Humans and other animals often learn about causal interactions between stimuli, actions, and outcomes with limited to no prior experience. Although humans may receive verbal or written instructions about the "rules," this type of explicit scaffolding is rarely available to non-humans. In such impoverished conditions, strategies such as exploration and hypothesis testing are important for identifying the best approach to maximize the goal (e.g., reward) and to build an internal model about the environment. Behavioral strategies that incorporate exploration can be executed randomly (no underlying pattern to the choices) or in a structured manner (directed exploration of particular choice options). Humans, including children and infants, often direct their exploration toward uncertain options or expectation violations[1–6] to seek information in an efficient way that aids learning.

Evidence for higher-order behavioral strategies is starting to emerge in rodent models of expert decision-making,[7–9] but the possibility that animals use such strategies during learning of a completely novel task has been more challenging to isolate. This remains true despite classic ideas around hypothesis testing in rodents first introduced by Krechevsky.[10] Such strategies typically come at a cost because animals defer immediate reward to learn about the causal structure of the environment.[11–13] As a result, systematic performance errors are challenging to dissociate from other sources of errors during learning. Common task designs in laboratory experiments require animals to learn that specific cues predict a desired outcome (e.g., reward) only after executing a given action (e.g., go or no-go or two alternative choices), giving rise to a multitude of error types related to sensory evidence, motor biases, or trial history that obfuscate the possibility of higher-order strategies. A challenge thus emerges in disentangling the nature and sources of various types of errors during learning.

Here, we exploit recent advances in computational modeling and experimental design to overcome this challenge. The development of a dynamic generalized linear model (GLM) to predict behavioral choices during learning[14] provides a tool to dissociate the temporally varying contribution of different factors during learning, including sensory evidence, motor bias, and trial history. In addition, recent work has demonstrated that the use of short blocks of non-reinforced "probe" trials can reveal learned

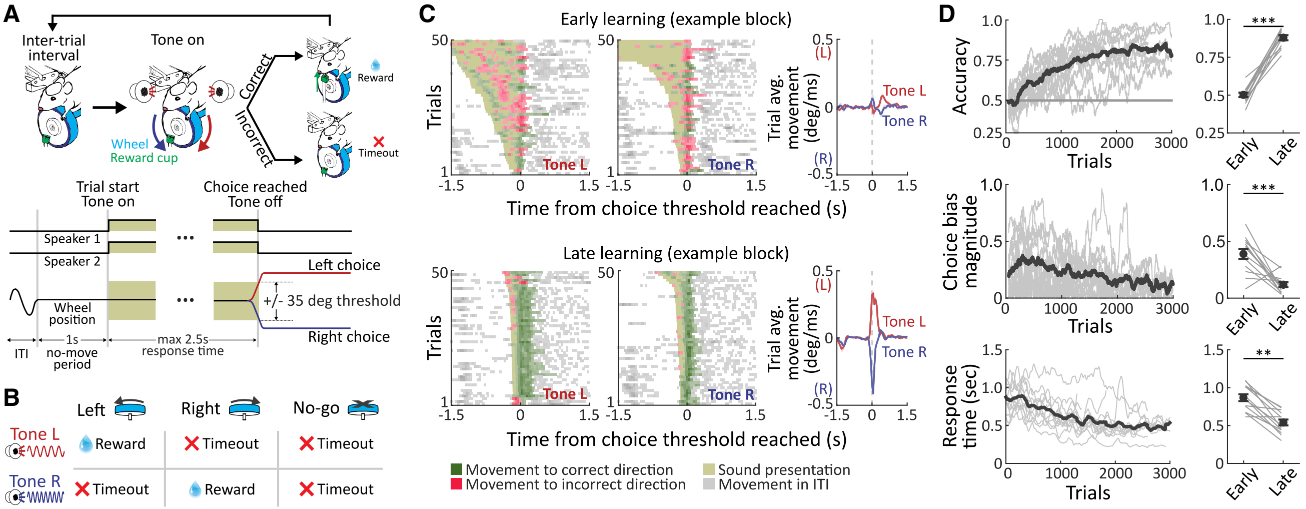


Figure 1. Learning of sensorimotor associations in an auditory two-choice task

(A) Schematic of the task. Top: schematic of the behavioral setup. Bottom: time structure of the task in one trial.

(B) Stimulus-action contingencies for the task.

(C) Wheel movement of an example mouse early and late in learning. Top left: wheel raster plot for early learning when animal's performance was at chance level, aligned to the time when choice threshold was reached, separated for left tone (tone L) and rightward-signaling tone (tone R). Trials are sorted by response time from slowest (top) to fastest (bottom). Green/red dots: movement toward the correct (green) and incorrect (red) direction during response window; gray dots: movement in the inter-trial-interval (ITI) that did not count toward the choice. The shade of the dots represents movement speed at each time bin of 50 ms, where darker shade indicates higher movement speed. Yellow shade: period of sound presentation. Top right: trial-averaged movement speed separated for different tones, represented as mean (solid line) ± SEM (shade). Red: movement speed on "tone L" trials, blue: movement speed on "tone R" trials. Bottom: similar with top panels, but for late learning when the animal reached expert-level performance.

(D) Trial-by-trial behavioral measurements during learning. Left: evolution of accuracy, choice bias, and response time during learning, calculated using 100-trial moving window for visualization purposes. Gray: individual animals, black: mean across all animals (n = 13). Right: accuracy, choice bias, and response time in early vs. late learning across animals. Gray: individual animals; black: mean ± SEM. Asterisks: * $0.01 \leq p < 0.05$, ** $0.001 \leq p < 0.01$, and *** $p < 0.001$.

See also Figure S1.

stimulus-action associations[15–17] and motor strategies,[18] independent of animal performance during task training. Finally, the development of wheel-based choice assays for rodents offers an opportunity to use an "analog" readout of choice to provide insight into an animal's decision process. We take advantage of all three tools to examine the strategies used by rodents when learning an auditory two-choice task from scratch.

RESULTS

Learning of sensorimotor associations in an auditory two-choice task

We trained head-fixed mice to perform a discriminative, wheel-based, two-choice task. Mice heard one of two continuous pure tones on a given trial and were rewarded for a correct left action (tone L) or right action (tone R). Incorrect actions or misses led to a short time-out (Figures 1A and 1B). Mice were required to turn the wheel 35° within a 2.5-s response window to reach a choice threshold and activate a lick-cup that moved upward to provide water (Figure 1A). To make the task more naturalistic, mice received binaural feedback coupled to the wheel movement, simulating a change of sound location in space as mice moved the wheel. Mice received pre-task shaping for up to 8 days to learn that turning the wheel left or right were the instrumental actions of interest. After this "wheel training," animals were put into the full, discriminative task with no additional shaping.

At the beginning of learning, mice moved the wheel sporadically after tone onset in both the correct and incorrect directions (example mouse, Figure 1C, "early learning," left panel); thus, the trial-averaged wheel movement following the onset of each stimulus is relatively flat (early learning, right panel). As training progressed, mice moved the wheel increasingly in the correct direction, time-locked to the tone onset with a much faster speed (Figure 1C, "late learning," left panel). The average wheel movement across each stimulus type thus diverges toward the direction of correct choice (late learning, right panel). Late in training, mice showed an increase in accuracy (early = 0.501, late = 0.879, p = 3.99e−09, n = 13 mice, Wilcoxon signed-rank test for all comparisons), reduction of choice-bias magnitude (calculated as absolute accuracy difference between left- and right-signaling trials, early = 0.389, late = 0.118, p = 6.16e−04, n = 13 mice), reduced response time (early = 0.868 s, late = 0.541 s, p = 1.40e−05, n = 13 mice, Figure 1D), increased wheel movement toward the correct side after sound onset (p = 0.03, n = 6 mice, Figure S1A), and increased action rate (defined as the proportion of trials with wheel movement above the choice threshold, either toward correct or incorrect direction) (p = 1.22e−03, n = 13 mice, Figure S1B), signaling faster, more accurate, and more vigorous decisions. Learning was also evident in wheel kinematics. Unsupervised clustering of trial-level wheel movements revealed four distinct clusters (Figures S1C–S1F): two corresponding to "fast" responses and the others corresponding to "slow" responses with delayed

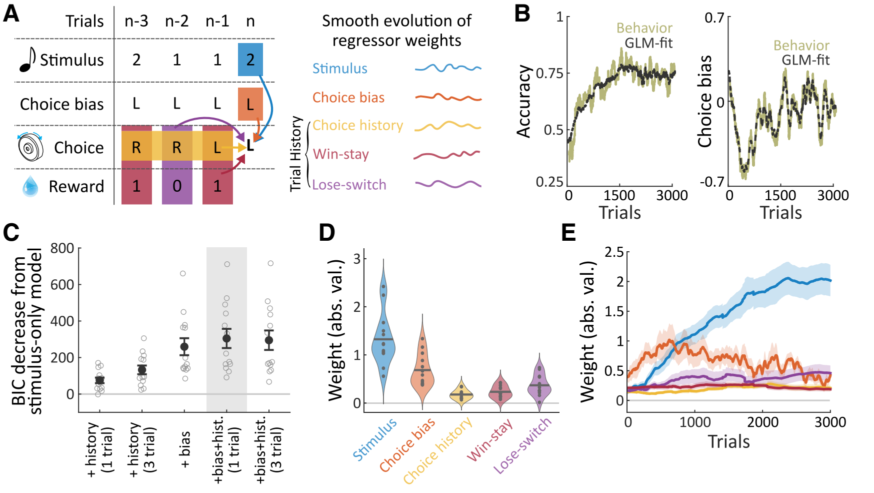


Figure 2. Choice bias is a major source of errors during learning

(A) Regressors used in the dynamic generalized linear model (PsyTrack GLM) used to predict choice probability on each trial.

(B) Model captures the accuracy (left) and choice bias (right) curve of individual mice during learning (example mouse). Experimentally observed accuracy and choice bias were calculated using a 100-trial moving bin. Green: behavioral data; black: accuracy and choice bias predicted by the PsyTrack GLM.

(C) Model selection by evaluating the decrease of BIC in various models compared with a base stimulus-only model. Gray: individual animals; black: mean ± SEM. Gray shade indicates that the "+bias+history (1-trial)" model (i.e. model with stimulus, bias, and one-trial history regressor) is the best model according to BIC.

(D) Distribution of average weight of each regressor during learning across animals, using the "+bias+history (1-trial) " model. Data are represented as a violin plot with mean (gray bar) to show the distribution of each regressor across animals with a Gaussian-mixture fit (color shades).

(E) Weight evolution of each regressor during learning, represented as mean (solid line) ± SEM (shade). Color represents the identity of the regressor as in (D).

See also Figure S2.

onset or back-and-forth wheel movements. As learning progressed, mice exhibited fewer trials with slow clusters, supporting the observation that response time was reduced as mice learned the task (Figures S1E and S1F).

Choice bias is a major source of errors during learning

We next sought to understand the drivers of errors during learning by expanding our analysis to the whole period of learning (using all trials from the start of learning to a maximum of five sessions at plateau level of performance). Errors could result from incomplete knowledge about stimulus-response contingencies or could result from a multitude of biases in decision-making. Immediate trial history could bias choices depending on the choice or outcome of previous trials, including the tendency to repeat the choice of previous trials (choice history bias), repeat a rewarded choice ("win-stay" bias), and to switch away from a previously unrewarded choice ("lose-switch" bias). In addition, animals may exhibit sustained or dynamic choice biases: repeatedly selecting a particular choice regardless of stimulus identity or recent trial history.

To disambiguate the sources of errors, we built a set of regressors based on these five terms and fit a dynamic GLM (PsyTrack) to predict the animals' choices during the whole learning process.[14] We included stimulus identity (binarized as 1/−1 for left/right trials), choice bias (captured by a constant term), choice history, win-stay, and lose-switch tendencies (with values of 1/−1 or 0, indicating a left/right tendency or no tendency). We included up to three previous trials in the history bias regressors (Figure 2A). The model predicts animals' choices trial-by-trial using a weighted sum of these regressors, with a logistic function that predicts choice probability on each trial. Importantly, PsyTrack allows regressor weights to evolve dynamically and smoothly across trials following a Gaussian process, which enables it to capture how the contribution of each regressor evolves during learning. Due to this smoothness constraint, the model-inferred choice bias can capture a slowly evolving choice preference on a much longer timescale (of 100–300 trials, Figures S2A and S2B) compared with immediate choice history. Such longer timescale patterns do not arise from randomness in behavior because model weight estimations were not affected by randomness in a simulated agent with random-choice behavior (Figure S2C).

Together, the best model fit included regressors of stimulus identity, choice bias, and trial history of one immediate previous trial (Figure 2C). This model accurately reproduced the experimentally observed accuracy and choice bias during learning (Figure 2B, average R-squared of accuracy fit = 0.932, choice-bias fit = 0.920, n = 13 mice, behavioral accuracy and choice bias calculated using 100-trial moving window).

What types of errors underlie these incorrect choices? We started with a stimulus-only model, added individual regressors to the base model, and evaluated model fit using the Bayesian information criterion (BIC), where a decrease in BIC indicates better model fitting. Besides the base model with the stimulus regressor (which captures incorrect stimulus-response knowledge), we found that adding the choice-bias regressor improved model performance (Figure 2C, "+bias" model: mean BIC decrease = 259.4) more significantly compared with adding the trial history regressor of the past one or three trials ("+history [1-trial]": mean = 75.0; "+history [3-trial]": mean = 132.6). Compared with the "stimulus+bias" model, adding either history term (either one-trial or three-trial back) only modestly further improved the model ("+bias +history [1-trial]": mean = 304.6; "+bias +history [3-trial]": mean = 294.9 compared with +bias model, mean = 259.4).

The weights of these regressors showed a similar trend. Averaged weights over learning showed that choice bias contributed much more than trial history regressors (Figure 2D, mean of each regressor: stimulus = 1.318, choice bias = 0.678, choice history = 0.170, win-stay = 0.227, lose-switch = 0.362). The contribution of choice bias was the highest early in learning and gradually decreased as stimulus weight increased. History biases, on the other hand, maintained a lower weight throughout learning, with lose-switch biases being the more prominent term

(Figure 2E). Together, these results suggest that early in learning, choice bias is the main source of error besides incorrect task knowledge, with history biases making a meaningful, but smaller, contribution.

Choice bias reflects a dynamic behavioral strategy rather than a stereotyped motor bias

The nature of choice biases observed during learning has long been puzzling. Choice bias could arise from inherent biases, such as motor preferences, or alternatively, from dynamic strategies, such as the structured exploration of a particular choice option.[8] To answer this question, we aimed to identify the structure of choice biases during learning. An inherent bias predicts that animals exhibit a consistent tendency to move the wheel toward one side (either left or right), whereas a dynamic strategy predicts that choice bias could fluctuate between both sides (Figure 3A).

To capture fluctuations of choice bias on shorter timescales than allowed by the dynamic GLM, we first quantified choice bias (i.e., signed accuracy difference between left and right-signaling trials) from behavioral data using a 40-trial moving window. We then divided trials into left-biased (L), right-biased (R), or unbiased (U) epochs whenever choice bias crossed over from the left (+) to right (−) side and calculated the area-under-the-curve (AUC) of each epoch. Epochs that significantly deviate from a chance distribution of AUC (obtained with the same method from a simulated random-choice agent with zero bias) were identified as significant (Figures 3B and 3C). Across all animals, we identified a significant number of biased (L or R) and unbiased (U) epochs using this method (L: 61 epochs, R: 69 epochs, U: 115 epochs, total n = 245 epochs), where most bias transitions did not correlate with significant changes in stimulus probability or reward rate (221 of 232 transitions). We identified transitions from unbiased to biased epochs (L or R) (U to L: 49 transitions, U to R: 50 transitions, across n = 13 mice), and from biased to unbiased epochs (L to U: 47 transitions, R to U: 56 transitions), and even transitions directly from left-to-right or right-to-left biases (L to R: 12 transitions, R to L: 7 transitions) (Figure 3D). Nearly all mice exhibited biased epochs on both sides (Figure 3E, 12 out of 13 mice). Finally, we conducted the same set of analyses on choice biases identified by the dynamic GLM, which allowed us to understand the evolution of choice bias at a slower timescale. The results revealed a similar dynamic structure in choice bias where we could identify a multitude of transitions between biased and unbiased epochs (Figures S3A–S3C).

Could these dynamic fluctuations in choice structure arise from choice randomness? To test this, we simulated agents with a constant bias that matches the choice-bias magnitude of each mouse and evaluated their choice-bias structure. Considering that each mouse may have a "preferred choice," we split biased epochs into their preferred side and non-preferred side. The preferred side was already apparent during wheel training (Figures S3D and S3E) and manifested as a stable but low movement bias to a single side (Figures S3F and S3G). Indeed, across all mice, the constant-bias agents almost exclusively exhibited biased epochs on the preferred side, whereas mice in our task exhibited biased epochs on both sides (Figure 3F). Mice spent significantly more trials in biased epochs on the non-preferred side, and significantly fewer trials on the preferred side compared with the constant-bias agent (mean proportion of biased trials on the non-preferred side: constant-bias agent = 0.0023, mouse behavior = 0.140, p = 4.88e−04; on the preferred side: constant-bias agent = 0.878, mouse behavior = 0.518, p = 2.44e−04, n = 13 mice, Wilcoxon signed-rank test) (Figure 3G), validating that the structure of choice bias is, indeed, dynamic.

The dynamic choice bias evolves throughout learning

Performance lapses occur even when rodents reach expert performance on a given task, often marked by sudden shifts into biased states or changes in lapse rate.[8,19–21] These lapses are hypothesized to reflect continuous exploration, but the nature and evolution of such exploration have been more difficult to assess during learning itself. Thus, we next asked if the evolution of choice bias during learning, indeed, reflected this evolving strategy.

We first quantified the average choice-bias magnitude (i.e., absolute value of signed choice bias) of each epoch and the number of transitions between biased and unbiased epochs as a function of trials in training (Figure 4A). As learning progressed, the magnitude of bias in each epoch decreased (early learning = 0.397, late learning = 0.274, $p = 6.34 \times 10^{-4}$, paired t test, n = 13 mice for all comparisons, Figure 4B), but the number of transitions stayed unchanged (early = 3.35, late = 3.45 per 500 trials, p = 0.895, Figure 4C). In addition, mice spent less time in high bias epochs (early = 0.350, late = 0.059, $p = 7.74 \times 10^{-4}$), and more time in unbiased epochs (early = 0.323, late = 0.592, $p = 1.68 \times 10^{-3}$), but the time spent in low bias epochs persisted throughout learning (early = 0.225, late = 0.197, p = 0.570) (Figures 4D–4F). This trend was not due to variability in the rate of learning between animals: performing the same analysis after aligning the learning curve of each animal to a common accuracy threshold of 70% (Figures S4A–S4D) led to similar results (Figures S4E–S4H).

In addition, we further confirmed the existence of biased bouts of exploration even when animals reached expert performance. We used a GLM combined with a hidden Markov model (GLM-HMM) to discover fast changes in strategies.[22] The model predicts choices using GLMs with stimulus and choice bias as regressor and uses an HMM to alternate between these GLMs on a single-trial level to capture state-like transitions. The best model (selected using cross-validated BIC, Figure S4J) contained three HMM-states, including a right-biased state, a left-biased state, and an unbiased state (Figure S4I). Mice exhibited dynamic shifts in strategies, evident in shifts between these three HMM-states within and between training sessions (example sessions, Figure S4K), with biased states occurring in epochs of 10 to 150 trials (Figure S4L). These results were consistent with our observation and further confirmed the dynamic nature of choice bias during learning.

Together, these results suggest that as learning progresses, high biases become rare, but low bias epochs and the dynamic structure of bias persists. These structured changes in choice bias over time suggest that previous reports demonstrating exploration at expert and near-expert levels[8] are a continuation of a dynamic strategy that evolves during learning.

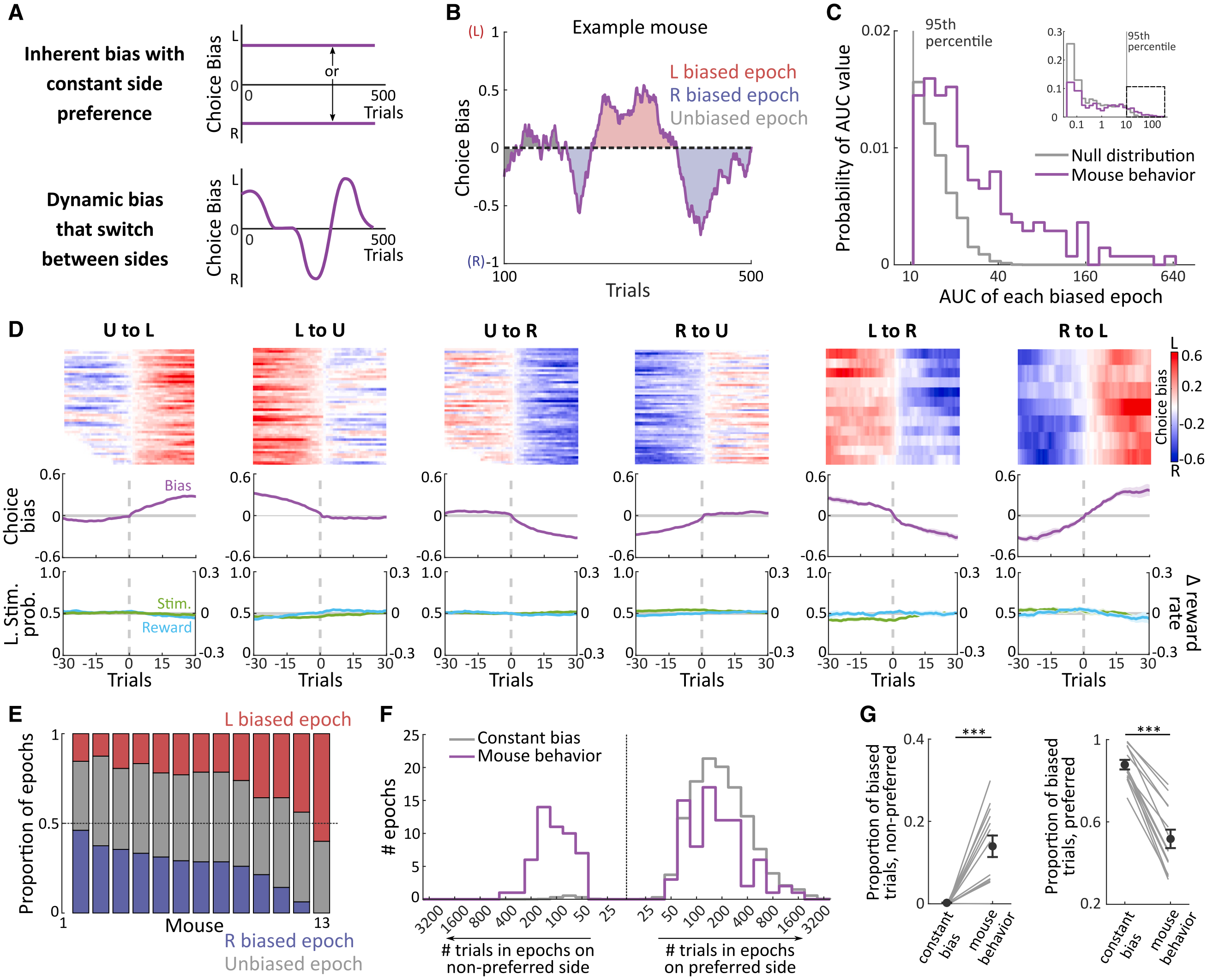


Figure 3. Choice bias reflects a dynamic behavioral strategy rather than a stereotyped motor bias

(A) Schematic of predicted choice bias by (1) an inherent bias that stays constant on either the left or right side or (2) a dynamic bias.

(B) Example of significant biased epochs identified from behaviorally quantified choice bias using a 40-trial moving window. Purple: choice bias; red/blue/gray shade: AUC of classified left-biased/right-biased/unbiased epochs.

(C) Probability distribution of AUC values over all epochs in all mice (actual behavior, purple) vs. chance distribution of AUC expected with a simulated random-choice agent with zero bias (gray), zoomed-in at the tail of distribution after 95th percentile of the chance distribution of AUC (the threshold for detection of biased epochs). Inset: the whole distribution of AUC values.

(D) Transitions identified between unbiased epochs (U) and biased epochs (L/R). Top: color plots showing choice bias (red/blue corresponding to left/right bias) during each transition. Bottom: choice bias (purple), stimulus probability (green), and change in reward rate (compared with pre-transition period, light blue) during epoch transitions, represented as mean (solid line) ± SEM (shade).

(E) Proportion of biased epochs for individual animals. Data are ordered by mouse based on the proportion of right-biased epochs, from high to low.

(F) Distribution of biased epoch length in behavioral data (purple) over all mice (n = 13), compared with simulated agents with constant biases that match the overall bias of each mouse (gray). Distribution is separated into epochs on the preferred or non-preferred side of each mouse's overall choice bias.

(G) Comparison between behavioral data and simulated agent. Left: proportion of trials in biased epochs on the non-preferred side. Right: same but for epochs on the preferred side. Gray: individual animals; black: mean ± SEM. Asterisks: * $0.01 \leq p < 0.05$, ** $0.001 \leq p < 0.01$; *** $p < 0.001$.

See also Figure S3.

Choice bias does not arise from decreased engagement

What could this dynamic structure of choice bias entail? Could it reflect a change in motivation or engagement,[20] or could it reflect a higher-order strategy? We analyzed the wheel kinematics to provide a richer understanding of the internal decision process of the animal. Specifically, we evaluated response latency (response time and initiation time), response vigor (corresponding to average movement speed), and movement during the inter-trial interval (ITI) (Figure 5A). In biased epochs, mice exhibited reduced latencies as bias magnitude increased (response time: r = −0.113, slope of linear regression = −0.086, p = 3.18e−03; initiation time: r = −0.119, slope of linear

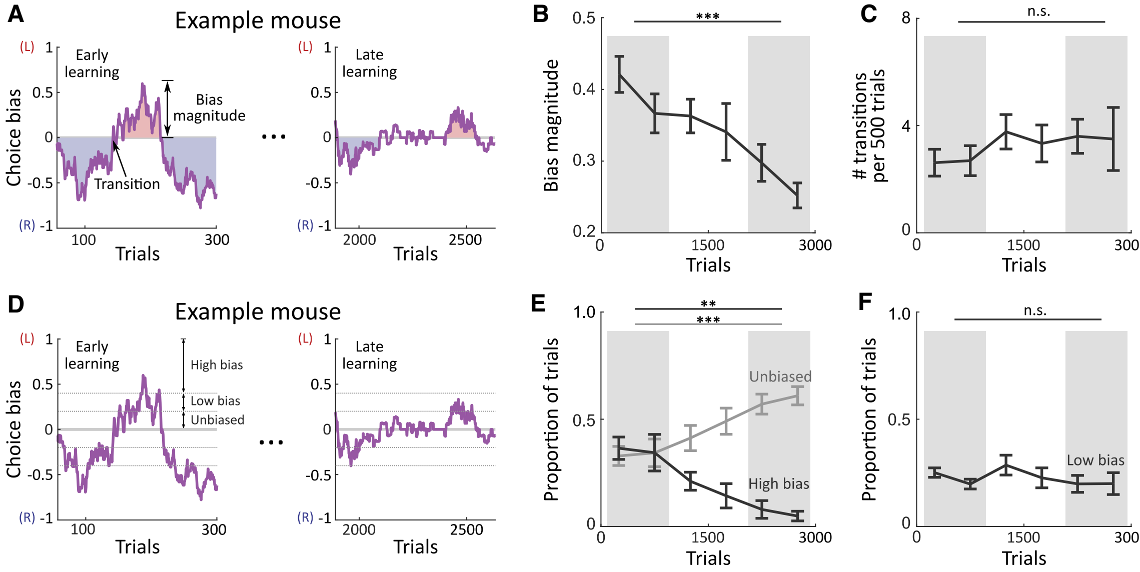


Figure 4. The dynamic choice bias evolves throughout learning

(A) Identifying bias magnitude of each epoch and the number of transitions between left/right-biased and unbiased epochs. Schematic of an example mouse.

(B) Plotting bias magnitude of each biased epoch as a function of number of trials in training (binned by 500 trials). Direct comparison was made between early learning (first 1,000 trials) and late learning (trial 2,000 to 3,000), indicated by the gray shaded area. Data plotted as mean ± SEM.

(C) Similar to (B) but plotting the number of transitions between left/right-biased and unbiased epochs as a function of trials in training. Data plotted as mean ± SEM.

(D) Classifying trials as high/low/unbiased trials using a threshold of choice-bias magnitude (low bias: 0.2–0.4, high bias: >0.4). Schematic of an example mouse.

(E) Similar to (B) but plotting the proportion of high biased (black line) or unbiased trials (gray line) as a function of trials in training. Data plotted as mean ± SEM.

(F) Similar to (B) but plotting the proportion of low-biased trials as a function of trials in training. Data plotted as mean ± SEM.

Asterisks (B)–(F): * $0.01 \leq p < 0.05$, ** $0.001 \leq p < 0.01$, and *** $p < 0.001$.

See also Figure S4.

regression = −0.074, p = 2.04e−03), while maintaining the same level of response vigor regardless of bias direction (the average movement speed was not correlated with bias magnitude, r = −6.68e−03, p = 0.86) (Figures 5B–5D). Decreased response latency was consistent with the observation that mice exhibited more trials within the fast cluster as bias increased (Figure S5B). Interestingly, mice also exhibited more movements toward the side of their choice bias during the ITI (r = 0.248, slope of linear regression = 7.24, p = 4.48e−06, Figure 5E). These correlations were only observed during biased epochs and were absent in unbiased epochs (Figure S5A).

If biased epochs result from disengagement, we would expect animals to miss more trials in biased epochs, and biased epochs would occur late in a training session when animals are more likely to disengage. However, we observed no difference in action rate (Figure S5C), and biased epochs occurred throughout the entire session (Figure S5D). In addition, the nature of the underlying driver of these biased epochs (strategy vs. disengagement) makes opposite predictions about learning during biased epochs. A dynamic strategy predicts that animals (1) learn faster with a higher bias magnitude, and (2) show greater learning during biased epochs compared with unbiased epochs. If biased epochs were due to disengagement, animals with more bias should learn slower, and biased epochs would show little-to-no learning. To investigate the effect of bias on the rate of learning, we used the dynamic GLM to estimate the stimulus-response knowledge of the animal by the predicted accuracy of the stimulus regressor. In support of a dynamic strategy, we observe that bias magnitude correlates with the rate of learning across animals (Figure S5E) and that there is a higher rate of learning in biased epochs compared with unbiased epochs (Figure S5F). Finally, disengagement would predict that animals pay less attention to the stimulus identity on each trial, which would lead to a decrease in accuracy on trials where animals chose the opposite side of their preferred choice in the current biased epoch. However, we observed an increase in accuracy on these trials (Figure S5G), suggesting that animals are still engaged and paying attention to the stimulus identity in these biased epochs.

Together, converging evidence from decreased response latency, maintained response vigor, and rate of learning suggests that biased epochs are unlikely to be periods of disengagement. Instead, they reflect a change in behavioral strategy that is apparent in wheel kinematics, both during the trial and in the ITI.

Choice bias reflects a strategy rather than incomplete knowledge about the task

If the dynamic choice bias is a strategy that supports learning, animals will have generated an expectation of reward on correct trials and no reward on incorrect trials. They should thus be sensitive to violations of that expectation (i.e., an unexpected change in the contingencies). Alternatively, if the dynamic choice bias is due to animals applying incorrect task knowledge, they will not be sensitive to such violations.

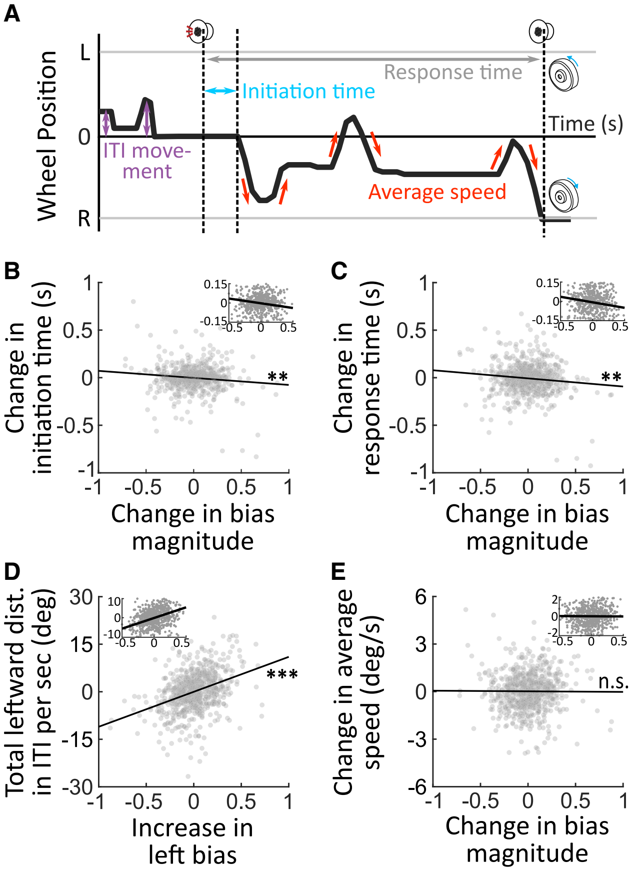


Figure 5. Choice bias does not arise from decreased engagement

(A) Attributes of the wheel trajectory reflecting response latency (response and initiation time), movement in the inter-trial-interval period (ITI movement), and response vigor (average speed).

(B) Correlation between changes in bias magnitude and wheel movement initiation time between two randomly selected periods of 40 trials within a biased epoch. Each gray dot represents the difference of bias magnitude and initiation time between two selected periods. Black line: line of linear regression of changes in bias magnitude and initiation time. Asterisks indicate the significance level of the correlation between the two terms. Inset: close-up scatter plot of choice-bias magnitude and initiation time between −0.5 and 0.5.

(C) Same with (B) but for correlation between changes in bias magnitude and response time. Inset: same with (B) but with response time.

(D) Same with (B) but for correlation between the directional increase in left bias and total leftward movement distance during ITI (positive: total movement is on left side, negative: on right side). Inset: same with (B) but with ITI movement.

(E) Same with (B) but for correlation between changes in bias magnitude and average wheel speed. Inset: same with (B) but with wheel speed.

Asterisks (B)–(E): * $0.01 \leq p < 0.05$, ** $0.001 \leq p < 0.01$, and *** $p < 0.001$.

See also Figure S5.

To test an animal's sensitivity to expectation violation, we utilized a probe trial design. During regular training with reinforcement ("reinforced" trials), we trigger a single catch trial where reward is omitted on a correct action (i.e., "catch-on-correct"), followed by a block of 10 continuously non-rewarded probe trials (Figure 6A). Reinforcement is reinitiated after the end of the probe block. This experimental design allows us to test the extent to which animals are sensitive to expectation violations by determining whether there was a significant change in choice behavior during the probe block.

Interestingly, mice exhibited strikingly different behavior during these probe blocks. Mice made many incorrect choices in reinforced trials but suddenly shifted to performing at higher accuracy and lower choice bias during the probe block after the single catch-on-correct trial (example mouse, Figure 6B). Specifically, this mouse made many errors by choosing the rightward action (red solid dots) on trials with a leftward-signaling tone (green hollow circle) in reinforced trials (gray shade), thus showing a choice bias to the right. In probe trials (yellow shade), however, the animal performed almost perfectly, with balanced choice on both left and right trials. Together, accuracy was much higher, and choice-bias magnitude was much lower during the probe trials.

To assess across animals and account for differences in their learning profile, we aligned the learning curve of each mouse to a 70% accuracy threshold in reinforced trials, which minimized the variance over their learning curves (Figures S4A–S4D). We focused our analysis on early learning (± 400 trials from performance threshold, 3–5 sessions). Across all mice, accuracy significantly increased (reinf = 0.681, probe = 0.781, p = 4.86e−04, Wilcoxon signed-rank test for all comparison, n = 13 mice) and bias magnitude significantly decreased (reinf = 0.328, probe = 0.238, p = 0.0217) in the probe block compared with reinforced trials (Figure 6C). Importantly, this effect was not driven by fluctuations within a session as it remained robust when we compared probe block performance with reinforced trials immediately before and after probe blocks (Figures S6A and S6B).

These observations reveal that even when mice have already established knowledge about stimulus-response contingencies, they trade off performance (and higher amounts of reward), potentially engaging in a dynamic choice-bias strategy to gain more information. This would predict that the difference in accuracy and bias between reinforced and probe would be greater if animals employ a higher biased strategy. Indeed, the degree of accuracy gain (Figure 6D) and bias reduction (Figure 6E) is dependent on session-level bias under reinforcement (accuracy: p = 0.0462, slope = 0.263, r = 0.317; bias magnitude: p = 6.67e−06, slope = −0.977, r = 0.646, linear regression, n = 40 sessions), which was not observed when trial identity was shuffled (accuracy: p = 0.736, slope = −7.74e−03, bias: p = 0.131, slope = 0.0630). In fact, choice bias in probe blocks remained at a low level regardless of choice bias in reinforced trials (Figure S6B). These correlation patterns remained robust when selecting sessions across the whole learning process (Figures S6C–S6E).

To test whether this correlation is unique to choice bias, we used the GLM to remove the choice bias or history regressors. Removing choice bias increased accuracy in a choice-bias-dependent manner (mean change = 0.038, p = 7.32e−04, n = 13 mice, one-sample Wilcoxon signed-rank test, r = 0.3474, p = 1.13e−16, slope = 0.148, linear regression), which is not observed when removing history terms (mean change = −9.4e−03, p = 0.0398, Wilcoxon signed-rank test; r = −0.0036, p = 0.934, slope = −9.31e−04, linear regression) (Figure 6F).

An alternative hypothesis is that the presence of reward in reinforced trials could drive more impulsive choices, a maladaptive state-dependent mechanism that arises from mice being over motivated. If this were the case, removing reward (in the

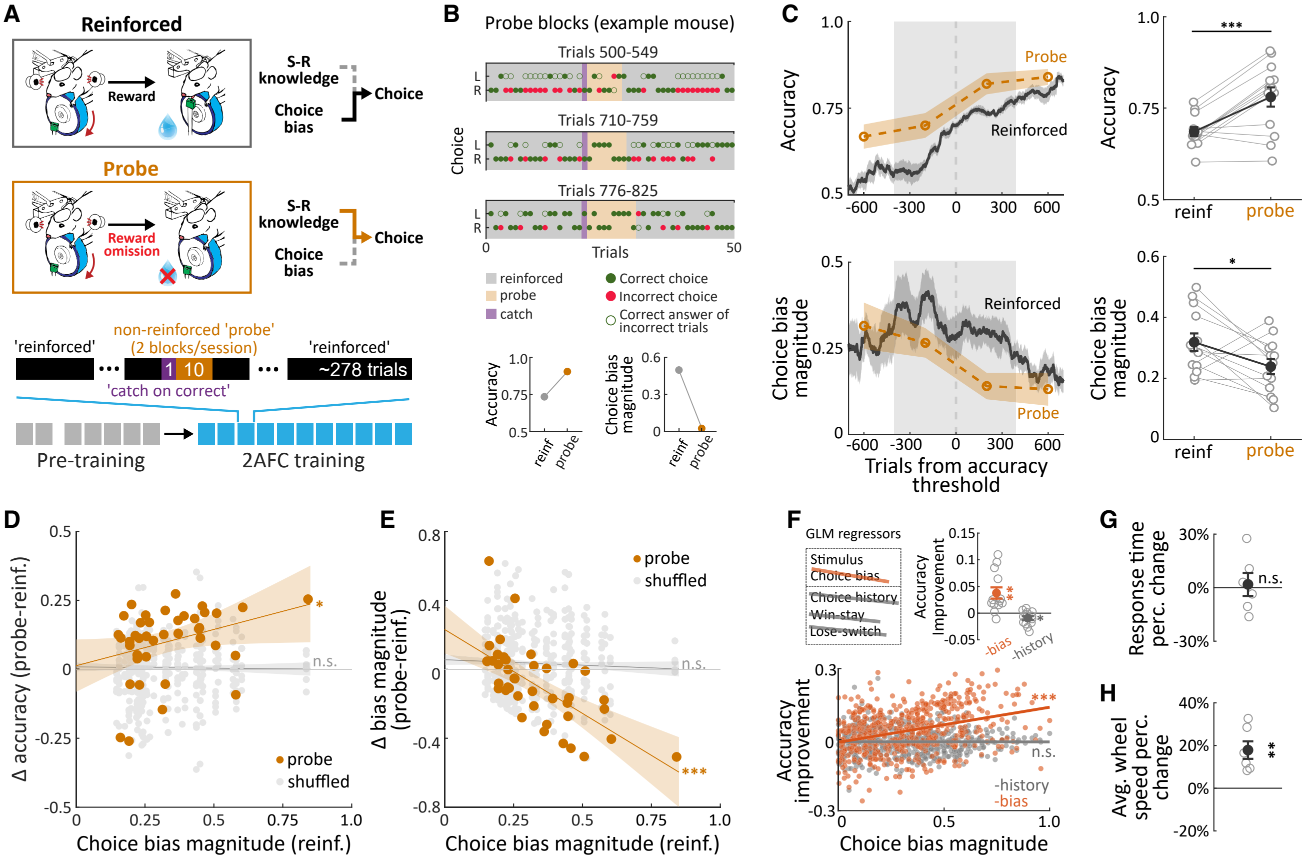


Figure 6. Choice bias reflects a strategy rather than incomplete knowledge about the task

(A) "Probe" trial design to test for knowledge of the task contingencies. Top: unexpected reward omission in probe trials was achieved by not moving the "lick-cup." Importantly, before the choice was made, the physical context was exactly the same between "reinforced" and probe trials (top left). We hypothesize that this unexpected omission shifts animals' choice toward testing stimulus-response knowledge (top right). Bottom: schematic of how two probe blocks are interleaved during each training session. Each block of probe trials is initiated by a "catch-on-correct" trial where reward is omitted on a trial where animals made a correct choice.

(B) Choices in probe blocks in an example mouse. Top: choices in three example probe blocks and the reinforced trials that preceded and followed. Green/red solid dots: correct/incorrect choice. Green hollow dots: the correct answer of incorrect trials. Gray/yellow/purple shade: reinforced/probe/catch trials. Bottom: averaged accuracy (reinf = 0.735, probe = 0.903) and choice-bias magnitude (reinf = 0.494, probe = 0.083) of example blocks in reinforced and probe trials.

(C) Comparison of accuracy and choice bias between reinforced and probe trials for all mice. Left: accuracy and choice-bias magnitude during learning (reinforced: 100-trial moving window, probe: 400-trial bin), aligned to an accuracy threshold of 0.7. Right: accuracy and choice-bias magnitude comparison in reinforced and probe trials, during ±400 trials of the accuracy threshold ($n = 13$ mice). Gray: individual animals; black: mean ± SEM.

(D) Relationship between choice-bias magnitude under reinforcement and change of accuracy in probe trials ($n = 40$ sessions). Orange/gray line: line of linear regression of probe/shuffled data. Orange/gray shade: indicating the 95% confidence interval of the predicted value of the linear regression model. Asterisks indicate the significance level of the correlation between the two terms.

(E) Similar to (D) but plotting the relationship between choice-bias magnitude under reinforcement and change of choice-bias magnitude in probe trials ($n = 40$ sessions) with linear regression analysis (legend same as in D).

(F) Removal of choice-bias or history-bias terms from the PsyTrack GLM. Top: schematic and average accuracy change upon removal across animals ($n = 13$ mice). Light gray: individual animals; dark orange/gray: mean ± SEM for removing choice-bias/history-bias terms. Bottom: correlation between accuracy improvement and experimentally observed choice bias, analyzed using 40-trial bins. Each dot represents data from one 40-trial bin. Dark orange/gray line: line of linear regression of removing choice bias/history biases from the PsyTrack GLM. Asterisks indicate the significance level of the correlation between the two terms.

(G) Percent change in response time in probe trials compared with reinforced trials across animals ($n = 6$ mice). Gray: individual animals; black: mean ± SEM.

(H) Percent change in average speed of wheel movement during the response period across animals ($n = 6$ mice). Gray: individual animals; black: mean ± SEM.

Asterisks (C)–(H): * $0.01 \le p < 0.05$, ** $0.001 \le p < 0.01$, and *** $p < 0.001$.

See also Figure S6.

probe block) should lead to decreased motivation, higher response latency, and reduced response vigor (i.e., slower wheel movements). However, we observed no significant difference in response time (mean percent change = 1.8%, $p = 0.789$, $n = 6$ mice, one-sample t test) and an increase in wheel movement speed that indicates increased vigor (mean percent change = 17.8%, $p = 0.0312$, $n = 6$ mice, one-sample Wilcoxon signed-rank test) (Figures 6G and 6H). These observations suggest that impulsivity is unlikely to drive differences in performance between reinforced and probe trials.

Together, these data demonstrate that mice are sensitive to expectation violations, abruptly shifting their behavioral strategy to test whether the outcome contingencies have changed.

DISCUSSION

Our results demonstrate that when mice learn a novel two-choice task with little-to-no previous experience, performance errors are dominated by a dynamic choice bias which, at first glance, appears maladaptive, but upon detailed analysis suggests that animals are guided by a higher-order behavioral strategy. We show that mice develop an internal model of the task such that violating their expectation of reward on correct trials (using a catch-on-correct reward omission trial followed by a short block of non-rewarded probe trials) leads to an abrupt shift in strategy. During the probe block, mice demonstrated an apparent knowledge of task contingencies with less bias compared with that in reinforced trials. This abrupt shift to using their acquired stimulus-action knowledge likely allows them to test whether the outcome contingencies have indeed changed. These observations suggest that on reinforced trials, rather than errors being driven by a lack of stimulus-response knowledge, they are instead engaged in a behavioral strategy that manifests as a choice bias. This strategy is dynamic in nature throughout learning, shifting between left, right, and unbiased epochs on a timescale of tens to hundreds of trials, co-occurring with reduced response latencies and selective movements in ITI. Although the magnitude of the bias decreases across learning, expert mice continue to exhibit short bouts of biases interspersed with longer bouts of unbiased choices and higher performance. These observations collectively suggest that choice bias during learning reflects a dynamic and continuous strategy that could serve a purpose during learning.

What could be the nature and purpose of this strategy? One hypothesis is that these strategies arise from information seeking through directed exploration. High uncertainty of the outcome of a given choice can direct animals' exploration to that choice,[1,4,9] thus giving rise to short bouts of biased choice patterns. This can produce a dynamic structure of biases consistent with our observation if the uncertainty over the two-choice options evolves dynamically. Such exploration also continues throughout learning, with a decreasing degree of exploration as the outcome of options becomes more certain. Alternatively, this biased strategy could benefit the animal by reducing cognitive effort. By preferentially sticking to one choice option, animals may need to spend less cognitive effort in making a choice and monitoring the outcome of that choice. Our data suggest that animals exhibit a combination of these two: they use alternating choice biases to engage in a lower-effort form of directed exploration.

In addition, our study adds to an increasing amount of literature which suggests that animals need to balance immediate reward gain against a multitude of longer-term goals, including information gain from exploration,[7,22] balance of cognitive effort and computational resources,[23–25] need of security,[26,27] and long-term behavioral flexibility.[28] These "strategic" goals are more challenging to isolate in laboratory settings where animals are trained, nominally, to maximize reward. Here, we demonstrate that by "probing" animals' stimulus-response knowledge—using the counter-intuitive and powerful approach of withholding reward—we could dissociate the goal of reward maximization (i.e., acting according to knowledge) from other behavioral goals, making the nature and form of individual learning strategies accessible.

This behavioral paradigm and analysis approach will also enable exploration of the underlying neural mechanism: what is the neural basis that governs individualized learning strategies, including the use of an alternating choice bias? What governs the balance between reward maximization (using stimulus-response contingencies to gain reward) and exploratory strategies? We hypothesize that the generation of these choice biases relies on a network of secondary motor cortex (M2) and posterior parietal cortex (PPC) since both areas are thought to be involved in bias generation.[8,29–35] In addition, the balance between adopting a biased strategy vs. acting according to task knowledge might be controlled by a higher-order structure, such as the pre-frontal cortex, which can integrate immediate needs with longer-term goals. Pre-frontal regions, including the orbital frontal cortex (OFC),[36,37] anterior cingulate cortex (ACC),[38,39] and frontal polar cortex,[40] have been shown to be important for driving exploration in human and non-human primates and thus could facilitate this choice-bias strategy as a means of exploring a given choice option.

Although we observe a dynamic structure of choice bias, it is important to note that animals still exhibit inherent biases to one side or the other. Mice perseverated more on a preferred side compared with the un-preferred side during the task (Figure 3F), and this inherent side bias was already apparent during pre-task wheel training (Figures S3D–S3G). Our data strongly suggest that inherent and dynamic biases can, and do, co-exist.

In summary, by combining non-reinforced probe blocks with computational approaches to understanding behavior, our study provides evidence that periods of low performance even at early stages of learning, likely reflect behavioral strategies where animals actively probe their environment in a structured manner.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Behavioral control systems
  - ○ Behavioral training and probe trial implementation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Measurements of behavioral performance and wheel kinematics
  - ○ Dynamic logistic regression model (PsyTrack)
  - ○ Identification and analysis of biased epochs
  - ○ Fitting and analysis of GLM-HMM model
  - ○ Comparison between reinforced and probe trials

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2024.04.017.

ACKNOWLEDGMENTS

We thank Dr. C. Drieu and J. Garmon for technical assistance in building the behavioral training system. We thank Drs. D. Lee, C. Fetsch, C. Drieu and members of the Kuchibhotla Lab for thoughtful feedback on our manuscript. This work was supported by grants from the NIH R01DC018650 and R00DC015014 (K.V.K.) and a JHU (Johns Hopkins University) Kavli NDI Distinguished Graduate Fellowship (Z.Z.).

AUTHOR CONTRIBUTIONS

Conceptualization, Z.Z. and K.V.K.; experiments, Z.Z.; data analysis, Z.Z., writing—original draft, Z.Z. and K.V.K.; writing—review & editing, Z.Z. and K.V.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 12, 2023
Revised: February 10, 2024
Accepted: April 8, 2024
Published: April 26, 2024

REFERENCES

1. Gershman, S.J. (2019). Uncertainty and Exploration. Decision (Wash D. C.) *6*, 277–286. https://doi.org/10.1037/dec0000101.
2. Blanco, N.J., and Sloutsky, V.M. (2021). Systematic exploration and uncertainty dominate young children's choices. Dev. Sci. *24*, e13026. https://doi.org/10.1111/desc.13026.
3. Almeras, C., Chambon, V., and Wyart, V. (2022). Competing cognitive pressures on human exploration in the absence of trade-off with exploitation. Preprint at PsyArXiv. https://doi.org/10.31234/osf.io/9qpuz.
4. Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., and Cohen, J.D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. J. Exp. Psychol. Gen. *143*, 2074–2081. https://doi.org/10.1037/a0038199.
5. Stahl, A.E., and Feigenson, L. (2015). Cognitive development. Observing the unexpected enhances infants' learning and exploration. Science *348*, 91–94. https://doi.org/10.1126/science.aaa3799.
6. Liquin, E.G., and Gopnik, A. (2022). Children are more exploratory and learn more than adults in an approach-avoid task. Cognition *218*, 104940. https://doi.org/10.1016/j.cognition.2021.104940.
7. Rosenberg, M., Zhang, T., Perona, P., and Meister, M. (2021). Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. eLife *10*, e66175. https://doi.org/10.7554/eLife.66175.
8. Pisupati, S., Chartarifsky-Lynn, L., Khanal, A., and Churchland, A.K. (2021). Lapses in perceptual decisions reflect exploration. eLife *10*, e55490. https://doi.org/10.7554/eLife.55490.
9. Wang, S., Gerken, B., Wieland, J.R., Wilson, R.C., and Fellous, J.-M. (2023). The effects of time horizon and guided choices on explore-exploit decisions in rodents. Behav. Neurosci. *137*, 127–142. https://doi.org/10.1037/bne0000549.
10. Krechevsky, I. (1932). "Hypotheses" in rats. Psychol. Rev. *39*, 516–532. https://doi.org/10.1037/h0073500.
11. Mehlhorn, K., Newell, B.R., Todd, P.M., Lee, M.D., Morgan, K., Braithwaite, V.A., Hausmann, D., Fiedler, K., and Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. Decision *2*, 191–215. https://doi.org/10.1037/dec0000033.
12. Cohen, J.D., McClure, S.M., and Yu, A.J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos. Trans. R. Soc. Lond. B Biol. Sci. *362*, 933–942. https://doi.org/10.1098/rstb.2007.2098.
13. Hills, T.T., Todd, P.M., Lazer, D., Redish, A.D., and Couzin, I.D.; Cognitive Search Research Group (2015). Exploration versus exploitation in space, mind, and society. Trends Cogn. Sci. *19*, 46–54. https://doi.org/10.1016/j.tics.2014.10.004.
14. Roy, N.A., Bak, J.H., International Brain Laboratory, Akrami, A., Brody, C.D., and Pillow, J.W. (2021). Extracting the dynamics of behavior in sensory decision-making experiments. Neuron *109*, 597–610.e6. https://doi.org/10.1016/j.neuron.2020.12.004.
15. Kurtenbach, H., Ort, E., Froböse, M.I., and Jocham, G. (2022). Removal of reinforcement improves instrumental performance in humans by decreasing a general action bias rather than unmasking learnt associations. PLoS Comput. Biol. *18*, e1010201. https://doi.org/10.1371/journal.pcbi.1010201.
16. Kuchibhotla, K.V., Hindmarsh Sten, T., Papadoyannis, E.S., Elnozahy, S., Fogelson, K.A., Kumar, R., Boubenec, Y., Holland, P.C., Ostojic, S., and Froemke, R.C. (2019). Dissociating task acquisition from expression during learning reveals latent knowledge. Nat. Commun. *10*, 2151. https://doi.org/10.1038/s41467-019-10089-0.
17. Oesch, L.T., Ryan, M.B., and Churchland, A.K. (2024). From innate to instructed: A new look at perceptual decision-making. Curr. Opin. Neurobiol. *86*, 102871. https://doi.org/10.1016/j.conb.2024.102871.
18. Mosberger, A.C., Sibener, L.J., Chen, T.X., Rodrigues, H., Hormigo, R., Ingram, J.N., Athalye, V.R., Tabachnik, T., Wolpert, D.M., Murray, J.M., and Costa, R.M. (2023). Exploration biases how forelimb reaches to a spatial target are learned. Preprint at bioRxiv. https://doi.org/10.1101/2023.05.08.539291.
19. Ashwood, Z.C., Roy, N.A., Stone, I.R., International Brain Laboratory, Urai, A.E., Churchland, A.K., Pouget, A., and Pillow, J.W. (2022). Mice alternate between discrete strategies during perceptual decision-making. Nat. Neurosci. *25*, 201–212. https://doi.org/10.1038/s41593-021-01007-z.
20. Hulsey, D., Zumwalt, K., Mazzucato, L., McCormick, D.A., and Jaramillo, S. (2023). Decision-making dynamics are predicted by arousal and uninstructed movements. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.02.530651.
21. Gupta, D., DePasquale, B., Kopec, C.D., and Brody, C.D. (2024). Trial-history biases in evidence accumulation can give rise to apparent lapses in decision-making. Nat. Commun. *15*, 662. https://doi.org/10.1038/s41467-024-44880-5.
22. Ashwood, Z., Jha, A., and Pillow, J.W. (2022). Dynamic Inverse Reinforcement Learning for Characterizing Animal Behavior. Proceeding of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022, *35*, pp. 29663–29676.
23. Kool, W., and Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. J. Exp. Psychol. Gen. *143*, 131–141. https://doi.org/10.1037/a0031048.
24. McGuire, J.T., and Botvinick, M.M. (2010). Prefrontal cortex, cognitive control, and the registration of decision costs. Proc. Natl. Acad. Sci. USA *107*, 7922–7926. https://doi.org/10.1073/pnas.0910662107.
25. Lai, L., and Gershman, S.J. (2021). Policy compression: An information bottleneck in action selection. Psychology of learning and motivation, *74* (Academic Press), pp. 195–232. https://doi.org/10.1016/bs.plm.2021.02.004.
26. Thompson, S.M., Berkowitz, L.E., and Clark, B.J. (2018). Behavioral and neural subsystems of rodent exploration. Learn. Motiv. *61*, 3–15. https://doi.org/10.1016/j.lmot.2017.03.009.
27. Whishaw, I.Q., Gharbawie, O.A., Clark, B.J., and Lehmann, H. (2006). The exploratory behavior of rats in an open environment optimizes security. Behav. Brain Res. *171*, 230–239. https://doi.org/10.1016/j.bbr.2006.03.037.
28. Molano-Mazón, M., Shao, Y., Duque, D., Yang, G.R., Ostojic, S., and de la Rocha, J. (2023). Recurrent networks endowed with structural priors explain suboptimal animal behavior. Curr. Biol. *33*, 622–638.e7. https://doi.org/10.1016/j.cub.2022.12.044.

29. Akrami, A., Kopec, C.D., Diamond, M.E., and Brody, C.D. (2018). Posterior parietal cortex represents sensory history and mediates its effects on behaviour. Nature *554*, 368–372. https://doi.org/10.1038/nature25510.

30. Hwang, E.J., Link, T.D., Hu, Y.Y., Lu, S., Wang, E.H.-J., Lilascharoen, V., Aronson, S., O'Neil, K., Lim, B.K., and Komiyama, T. (2019). Corticostriatal flow of action selection bias. Neuron *104*, 1126–1140.e6. https://doi.org/10.1016/j.neuron.2019.09.028.

31. Hwang, E.J., Dahlen, J.E., Mukundan, M., and Komiyama, T. (2017). History-based action selection bias in posterior parietal cortex. Nat. Commun. *8*, 1242. https://doi.org/10.1038/s41467-017-01356-z.

32. Erlich, J.C., Brunton, B.W., Duan, C.A., Hanks, T.D., and Brody, C.D. (2015). Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. eLife *4*, e05457. https://doi.org/10.7554/eLife.05457.

33. Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature *520*, 220–223. https://doi.org/10.1038/nature14066.

34. Guo, Z.V., Li, N., Huber, D., Ophir, E., Gutnisky, D., Ting, J.T., Feng, G., and Svoboda, K. (2014). Flow of cortical activity underlying a tactile decision in mice. Neuron *81*, 179–194. https://doi.org/10.1016/j.neuron.2013.10.020.

35. Li, N., Chen, T.-W., Guo, Z.V., Gerfen, C.R., and Svoboda, K. (2015). A motor cortex circuit for motor planning and movement. Nature *519*, 51–56. https://doi.org/10.1038/nature14178.

36. Johnson, C.M., Peckler, H., Tai, L.-H., and Wilbrecht, L. (2016). Rule learning enhances structural plasticity of long-range axons in frontal cortex. Nat. Commun. *7*, 10785. https://doi.org/10.1038/ncomms10785.

37. Schreiner, D.C., and Gremel, C.M. (2018). Orbital frontal cortex projections to secondary motor cortex mediate exploitation of learned rules. Sci. Rep. *8*, 10979. https://doi.org/10.1038/s41598-018-29285-x.

38. White, J.K., Bromberg-Martin, E.S., Heilbronner, S.R., Zhang, K., Pai, J., Haber, S.N., and Monosov, I.E. (2019). A neural network for information seeking. Nat. Commun. *10*, 5168. https://doi.org/10.1038/s41467-019-13135-z.

39. Jahn, C.I., Grohn, J., Cuell, S., Emberton, A., Bouret, S., Walton, M.E., Kolling, N., and Sallet, J. (2022). Strategic exploration in the macaque's prefrontal cortex. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.11.491468.

40. Zajkowski, W.K., Kossut, M., and Wilson, R.C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. eLife *6*, e27430. https://doi.org/10.7554/eLife.27430.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Isoflurane	MWI Animal Health	CHEBI: 6015
Deposited data		
Raw and analyzed data (available upon request)	This paper	N/A
Experimental models: Organisms/strains		
C57BL/6J mouse	Jackson Laboratory	RRID: IMSR_JAX: 000664
ChAT-IRES-Cre mouse	Jackson Laboratory	RRID: IMSR_JAX: 006410
Software and algorithms		
MATLAB	Mathworks	RRID:SCR_001622
Python	https://www.python.org	RRID:SCR_008394
Psytrack toolbox	Nicholas Roy (Jonathan Pillow Lab)	https://github.com/nicholas-roy/psytrack
SSM toolbox for GLM-HMM model fitting	Scott Linderman lab	https://github.com/lindermanlab/ssm
Code for analysis and data visualization	This paper	https://github.com/thekuchibhotlalab/analyze2AFC
Arduino IDE	https://www.arduino.cc/	RRID:SCR_024884
Behavioral control interface	Coulbourn Instruments	Graphic State 4
Other		
Electrostatic Speakers	Tucker-Davis Technologies	ES1
Speaker Driver	Tucker-Davis Technologies	ED1
Behavioral control system	Coulbourn Instruments	LabLinc H02-08
Arduino UNO	https://www.arduino.cc/	RRID:SCR_017284
Teensy microcontroller	https://www.pjrc.com/teensy/	Teensy 4.1
Rotary encoder	Autonics	E20HB3-360-3-N-5-S
Pneumatic cylinders	Fabco-Air	H-5-O-NR
Manual micromanipulator	Siskiyou	DT-100

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Kishore Kuchibhotla.

Materials availability

Custom-designed behavioral setup and 3D models of individual parts are available upon request.

Data and code availability

All original code for analysis and reproducing the figures in the paper has been deposited at Github (at https://github.com/thekuchibhotlalab/analyze2AFC) and is publicly available as of the date of publication. Processed behavioral data (in MATLAB data format) is also available upon request. Accession numbers are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All animal procedures were approved under IACUC protocol at Johns Hopkins University. Mice were housed in cages of 2-5 mice on a 12hr-12hr inverted day-night cycle under temperature (70-74 degrees Fahrenheit) and humidity control (40-60% humidity). Mouse housed in the same cage were same sex littermates. Male and female adult mice of ages from 8–16 weeks (at the start of behavior) were used (male n=12, female n=1 are included in the analysis). Mice used were C57 background, including C57BL/6J wildtype mice (Jackson Laboratories, n=12) and ChAT-cre strain (Jackson Laboratories, n=1).

METHOD DETAILS

Behavioral control systems

All behavioral events (stimulus delivery, reward delivery, inter-trial-intervals) were monitored and controlled by Lab Linc systems (Coulbourn Instruments). The auditory stimuli were generated by Teensy 4.1 microcontrollers with audio shield, amplified and delivered through ED1 speaker driver and ES1 speakers (Tucker-Davis Technologies). Water was delivered using a custom-made water cup attached to an air cylinder (Fabco-Air), and licks were detected using an infrared beam and detectors. The wheel was made to be 1.5 inch in diameter with silicon foam attached on the edge. The mouse tube, head-fixation system and the wheel were custom made. Wheel positions were recorded with rotary encoders with either 0.36 or 1-degree precision (Autonics). The position of head-fixation clamp and the wheel was adjusted using manual precision micro-manipulators (Siskiyou).

Behavioral training and probe trial implementation

After head-plate surgery, mice were given at least 7 days of rest until their weight stabilized before starting water restriction. Mice were water restricted for at least 5 days before behavioral training, where their weight was kept between 80-85% of their original weight. During behavioral sessions, mice sit on an acrylic tube, with their forearms placed on wheel. Mice were first habituated to head fixation for 3 days, and then went through pretraining consisting of 2-3 (median = 2) days of un-cued lick training (lick to receive a water reward), and 2-5 days (median = 5) of un-cued wheel training (move the wheel to reach a threshold on left or right to receive a water reward). Water reward was 2.5uL or 4uL depending on the animal. Wheel was fixed during habituation and lick training but allowed to move starting from wheel training. In wheel training, we count a wheel movement as a ‘significant wheel movement’ only if it reached a threshold of a certain degree (varied from 11 degrees in the beginning of wheel training to 35 degrees at the end of wheel training). If a continuous movement stopped for more than 150ms, we re-started counting the movement from 0. The goal of the wheel training was to shape the animals’ behavior by pairing left and right wheel movements with reward, and to reduce animals’ innate biases by rewarding left and right movements in a balanced fashion. Thus, there were no external cues that determined the delivery of reward, and animal were rewarded after a wheel movement that reached movement threshold, only when the following conditions were met. (1) A certain amount of time had passed after the last reward. This time-out period was increased from a fixed 2 seconds to a uniform distribution with a mean of 4 seconds (range from 2.5 to 5.5 seconds) at the end of wheel training. (2) The total number of rewards animals received after left and right movement should not have a difference greater than a specific number. For example, in the beginning of wheel training, we set this difference threshold to 5 rewards. If an animal has already gotten 50 rewards on left movements and 55 rewards on right movements, the next reward delivery would happen only when animal makes another left movement, instead of right. After that, the count becomes 51 for left and 55 for right, and the animal can receive a reward following the next movement, no matter if it is left or right. This means that animals will not be continuously rewarded if they consecutively move to the same direction above a certain number of trials. They need to alternate the direction of movement to get continuously rewarded. This difference threshold between left and right-side rewards were reduced from 5 at beginning of wheel training to 1 at the end of wheel training to encourage animals to switch between left and right movements. At the end of wheel training, mice were motivated and actively exploring movements on both directions, achieving 100-150 rewards on both leftward and rightward movements within 40 minutes.

After pretraining, mice were directly trained on the complete 2AFC task. In this task, mice were trained to discriminate between two pure-tone stimuli that were $^3/_4$ octaves apart. One cohort was trained on a pair of pure tones of 6.2 vs. 10.4kHz, and another group on 7 vs. 11.7kHz. The sound was delivered through two speakers on either side of the mouse, equal in distance. One tone signals a left choice being rewarded, and the other signals the right choice being rewarded. Task contingency is counterbalanced across animals. At the start of each trial, a continuous tone was played, and mice needed to turn the wheel to a left or rightward choice threshold to indicate a left or right choice within a 2.5 second response window. The threshold was set to 23 degrees in the initial 2-3 days of training to help mice reach choice threshold, and then set to 35 degrees afterwards. To make the task more natural to the animal, we implemented binaural feedback where the intensity of each sound was coupled to the wheel position. The initial intensity of both speakers was 65dB (~15dB above noise level). As mice move the wheel towards one direction, the volume on the contralateral side increased and the volume on ipsilateral decreased proportionally to the amount of movement, simulating a moving sound source. At the choice threshold, the contralateral speaker volume would increase by 5dB and ipsilateral speaker would be silent. The volume did not change further after the choice threshold was reached. The sound terminated immediately if the mice reached choice threshold or terminated at the end of response window if the mice did not reach threshold within the window (miss trials). A correct trial was rewarded with a 4uL of water reward (2 seconds consummatory period + 2 second delay), whereas incorrect or miss trials were followed by a timeout of 4 or 7 seconds (depending on the animal). ITI with a one-second no-move interval was implemented where the mice needed to hold the wheel relatively stable (making a total movement of less than 23 degrees) for 1 second before the start of the next trial.

Each day, mice were trained for a total of 300 trials with only a few exceptions. Within these 300 trials, we interleaved two blocks of 10 non-rewarded ‘probe’ trials (20 trials total, occasionally 1 block of 20 trials). ‘Probe’ trial blocks were at least 60 trials apart from each other and 80 trials apart from the beginning of the session. ‘Probe’ trials were always triggered by a ‘catch’ trial where, after mice made a correct choice on a trial, reward was omitted and the reward port remained unavailable, just as an incorrect trial. A ‘probe’ trial block immediately followed ‘catch’ trials where rewards were omitted every trial, regardless of whether animal’s choice was correct or not, using the same method as ‘catch’ trials. After ‘probe’ trial block ended, mice immediately transitioned back into reinforced trials where they would be rewarded if a correct choice was made.

QUANTIFICATION AND STATISTICAL ANALYSIS

Measurements of behavioral performance and wheel kinematics

We analyzed the whole period of learning by using all trials from the start of learning to max five days at plateau level of performance. All analysis in Figures 1–5 (unless specified) are conducted over all trials over learning. Accuracy, choice bias, action rate and reaction time were computed using a moving window of 100 trials throughout learning. Miss trials were removed from all analysis including trial number calculation, except action rate. Choice bias was calculated as the signed difference between accuracy on left and right stimuli (positive: left bias; negative: right bias). Choice bias magnitude was defined as the absolute value of choice bias, representing the magnitude of this bias regardless of direction preference. Wheel position was first pre-processed to 1ms-precision, from which movement speed was calculated. Wheel movement was then computed using 20ms bin for visualization purposes. All analysis was conducted from the start of learning to maximum five days at plateau level of performance. To compare behavioral metrics between early and late learning, we selected trials in early/late learning for each animal depending on their performance. For accuracy, response time, and action rate, we selected the first 300 trials as ‘early learning’ and the consecutive 300 trials of best accuracy at the end of learning as ‘late learning’. For choice bias magnitude, we consider that choice bias can dynamically fluctuate, and thus taking a fixed period may not accurately capture the magnitude. We identified the period of 300 trials with max choice bias magnitude in the first 1000 trials of training (as choice bias magnitude ‘early’ in learning), to compare with ‘late’ learning, where choice bias magnitude is quantified in the same window as other behavioral metrics.

Analysis of detailed wheel kinematics was performed in six mice for which precise wheel kinematics were recorded at millisecond precision. On each trial, we first focused on wheel kinematics during the response window (i.e. from tone onset to the time when choice threshold is reached). We extracted six attributes: initiation time, initiation speed, initiation distance, total time, total distance, and total time. We defined the ‘initial movement’ as the first continuous movement that exceeded 7 degrees after the tone onset and 150ms dead period. This was then used to calculate the time of initiation (initiation time), speed of the initial movement (initiation speed) and the distance of initial movement (initiation distance). We then looked at all movements during the response period (before response threshold is reached) and computed the time between the first movement and decision time (total time), the distance of all movements (total distance) and the average speed of all movements (total speed). Initiation and total movement were normalized to the choice threshold. We concatenated wheel trajectories of all trials and performed K-means clustering using these six attributes. We quantified the within-cluster sum of squared errors (WSS) and identified the appropriate number of clusters by identifying the elbow point of WSS. In addition to clustering analysis, we also analyzed spontaneous movements in ITI which were computed as the total wheel position change in the 4 second period before the start of each trial.

Dynamic logistic regression model (PsyTrack)

PsyTrack, a dynamic generalized linear model (GLM) with a logit link function, was fit to predict individual animal’s choices trial-by-trial during learning.[14] The stimulus regressor was implemented as +1 or -1 for leftward/rightward signaling stimulus. Choice bias regressor was implemented as a constant of +1 on all trials, whereas choice history regressor as the choice (+1 or -1) of the nth previous trial. Win-stay regressor was only implemented on previously rewarded trials (by putting the regressor to 0 on unrewarded trials) and lose-switch regressor only on previously un-rewarded trials (with probe blocks being excluded). Weights of each regressor were allowed to fluctuate smoothly under a gaussian process, and the smoothness was determined by a smoothness hyperparameter. Another hyperparameter governed abrupt changes in weight across training sessions. Hyperparameters were fit using maximum likelihood with 5-fold cross-validation, and the weight was inferred using maximum likelihood estimation.[14] The best GLM model was selected as the model of the lowest BIC.

GLM’s predicted accuracy on this task was calculated as the probability of the correct choice as predicted by the GLM. Predicted choice bias of the GLM was calculated as the difference of GLM predicted accuracy between left and rightward signaling trials, using a 100-trial moving bin. The fitness of the GLM on the behavioral learning curve was calculated as the correlation between GLM and behavioral accuracy/choice bias. To evaluate the contribution of weights, we took the absolute value of weights of all regressor. We restricted weight averaging from the onset of learning to when animals reached their peak performance. To visualize regressor weight during learning, we fixed the regressor weights if the animal has reached expert level before 3000 trials of learning. To assess the effect of removal of choice bias or history regressors, we started with ‘stimulus + choice bias + history (1-trial)’ model, set either choice bias weight to 0, or the weights of all history terms to 0, and then evaluate the change in GLM’s predicted accuracy. Finally, we also computed the GLM-inferred accuracy from stimulus weight (Figure S5E), which is different from GLM’s predicted accuracy on the task. GLM-inferred accuracy from stimulus weight was computed by setting all bias weights in GLM as 0, and only keeping the inferred stimulus weight. We then evaluate the accuracy of this stimulus-only model. This accuracy measures the animal’s knowledge about stimulus-action associations, as inferred by the GLM.

To understand the relationship between model-predicted choice bias and behaviorally measured choice bias, we evaluated the correlation of GLM-predicted choice bias and behavioral choice bias inferred at different lengths of a moving window. The moving window length that gave the best correlation reflected the ‘smoothness’ of the GLM estimation. To further confirm that evolution of weight is not influenced by random noise, we simulated a random choice agent that makes completely random choices regardless of stimulus type. The model with stimulus and choice bias were fit to the choices of random choice agents, while behavioral accuracy and choice bias measurements were calculated (same methods with previous section) using 100-trial moving window.

Identification and analysis of biased epochs

We identified biased epochs in two ways. First, we used behavioral choice bias to identify fast transitions in choice bias structure. Choice bias is calculated using a 40-trial bin as the difference of accuracy between left and right signaling trials. To identify significant transitions, we divided choice bias into epochs whenever it crossed over zero and computed the area-under-the-curve (AUC) of each epoch. We simulated a random choice agent with no choice bias and quantified the distribution of AUC values in this agent. The 95th-percentile of this chance distribution is set as the threshold, and significant bias epochs in behavior are thus defined by epochs with AUC greater than this threshold. Epochs that were not significantly left or right biased were classified as unbiased epochs. Unbiased epochs that were smaller than 10 trials were combined into the nearest biased epochs. Then, we also identified choice bias epochs using GLM-predicted choice bias. Since such choice bias estimate is not subjective to fluctuations, we directly inferred bias epochs whenever this choice bias estimate crosses over zero. We inferred 'unbiased' epochs as ones where the maximum expected choice bias is smaller than 0.2.

We simulated fixed bias agents based on the logistic choice regression model of each mouse. For each mouse, we took the fitted GLM model (stimulus and choice bias) and computed the average weight amplitude of choice bias (taken as the absolute value). We then simulated a fixed-bias agent whose choice bias matches the average amplitude of choice bias, but only stays on one side (i.e. always positive for left bias or negative for right bias). We simulated trial-by-trial choices of this agent, computed behaviorally inferred choice bias using a 40-trial bin, and identified significant bias epochs using the same AUC method.

To investigate how choice bias influences wheel kinematics within a trial and during inter-trial interval (ITI), we first focused our analysis on significant bias epochs. For each bias epoch, we divided it into bins using a 40-trial moving window that moves by 10 trials. For each behavioral measurement, we computed its change between two adjacent but non-overlapping 40-trial bins and built a linear regression model between change in bias magnitude (i.e., absolute value of choice bias) and changes in behavioral measurements (e.g., initiation time, response time, average speed and proportion of 'fast' clusters). The significance of the slope of linear regression was inferred by p-value of its t-statistics. For spontaneous movement in the ITI, since the direction of bias could influence the direction of ITI movements, we built a linear regression model between changes in leftward choice bias and the changes in total leftward direction (i.e., wheel position at the end of ITI - position at the start of ITI). We further conducted the same analysis on unbiased epochs as a control. Finally, we compared the learning rate, action rate, and accuracy of trials where animals chose the choice that is opposite of the 'preferred choice' between biased epochs and unbiased epochs. To make a pair comparison, for each transition between biased and unbiased epoch, we took 40 trials that happened chronologically together in each epoch, quantified these measurements, and averaged across all epoch transitions within each animal. These measurements were then compared across mice.

To evaluate the change in choice bias structure during learning, we binned the first 3000 trials of training into six 500-trial bins. For each bin, we identified all transitions between epochs, and identified 'biased trials' for all trials inside a bias epoch. In these 'biased trials', we computed mean bias magnitude as the average amplitude of choice bias, and we further divided them into 'high biased trials' (choice bias magnitude > 0.4), 'low biased trials' (choice bias magnitude < 0.4 and >0.2). 'Unbiased trials' were defined as trials either with <0.2 choice bias, or within an unbiased epoch. Same analyses were repeated by first aligning learning curves of individual animals (see section below) and then binning into 500 trial bins.

Fitting and analysis of GLM-HMM model

We aimed to fit a GLM-HMM model to trial-by-trial choice data as a model-based approach to test the existence of biased and unbiased epochs that we observed. While robust fitting during learning was challenging, we were able to fit the GLM-HMM model on choice data after animals reached expert-level.

Since we did not train animals on prolonged periods of expert-level performance, we concatenated choice data from all animals that reached expert-level performance to increase the robustness of model fitting. We fit a GLM-HMM using trial-by-trial choices of a total of 33 days, 8519 trials from 8 mice during expert-level performance. We defined 'expert performance' here as days where the PsyTrack GLM-inferred accuracy from stimulus weight reached 90% performance. Each state in HMM contains a Bernoulli GLM defined by a weight vector specifying how stimulus inputs (binarized as +1 and -1 for left/right) and choice bias (a constant of +1 representing a left-choice preference) are integrated in that state. The model was fitted using code and methods from previously published expectation-maximization (EM) algorithm.[19] To identify the optimal number of states, we evaluated the BIC using 5-fold cross-validated (i.e., splitting the data into training and testing in 4:1 ratio). A 3-state model achieved the best cross-validated BIC value to explain the choice behavior, capturing right-biased state, a left-biased state and unbiased state. To infer the most-likely HMM-state of each trial, we evaluated the state probability of each trial, and selected the state with the highest probability as the most-likely HMM-state. We further divided all the trials into epochs by which HMM-state the animals were in and quantified the length of each epoch.

Comparison between reinforced and probe trials

To compare reinforced and probe performance, we first defined the 'probe' trial of comparison. Since we always run 10 'probe' trials after a 'catch' trial, we selected those 10 trials for all probe comparisons. Then, we consider that, early in learning, animals may not have established any association between stimulus and the appropriate actions. Late in learning, repeated probing that violates animal's reward expectation may drive extinction in probe trials. Since animals learned at different rates, we first aligned the learning curve of all animals by the time they first reached a performance threshold under a 300-trial window to compute accuracy. We varied the accuracy threshold and selected the accuracy threshold = 0.7 since it best reduced the variance of learning curve across animals.

After aligning the learning curves, we examined performance on probe blocks 400 trials before and after reaching the defined accuracy threshold (total of 3-5 sessions of training, median = 4 sessions). Since animals' choice bias is dynamic and could change across days, we combined the two probe blocks in each day. We removed all 'miss' trials in the probe condition. Days containing less than 15 probe trials were combined into an adjacent day. For each session, we randomly subsampled the same number of reinforced trials, repeated 1000 times, to compute the expected accuracy and choice bias magnitude in reinforced trials.

For comparison of accuracy and choice bias magnitude, we combined multiple sessions of each animal; the same comparison was also conducted on session-by-session level, and on reinforced trials immediately before or after the probe block (where no sub-sampling was needed). To assess changes in behavioral measurements between reinforced and probe trials, we directly compared probe trials with reinforced trials immediately preceding them. Session-by-session data was also used for linear regression analysis. The 'shuffled' group in linear regression was generated by randomly sub-sampling two adjacent blocks of reinforced trials in each session (matching the trial number of probe trials) and computing the difference of accuracy or choice bias magnitude between these two groups.